# VALIDATING ACOUSTIC MEASURES OF SPEECH RHYTHM FOR SECOND LANGUAGE ACQUISITION

*Naoko Kinoshita & Chris Sheppard*

Meikai University, Japan; Waseda University, Japan
naoko-k@meikai.ac.jp; chris@waseda.jp

## ABSTRACT

This paper reports research investigating the validity of using Pairwise Variability Indexes in research into the second language acquisition of speech rhythm. Findings determined that 1) expert native-speakers rate non-native speaker rhythm based on a common factor, and 2) part of that common factor can be accounted for by the use of vocalic pairwise variability.

It was concluded that the PVI measures are indeed a valid measure of second language speech rhythm.

**Keywords:** language rhythm, PVI, L2 acquisition, native speaker evaluation

## 1. INTRODUCTION

The acquisition of second language speech rhythm is, demonstrably, a challenge for language learners [14]. A clear understanding of the processes underlying this acquisition should assist both language learners and teachers. Until recently, the method used to measure Japanese rhythm was a ratio measure which compares the lengths of special mora and non-special mora, i.e. Kashima [7]. However, measures like this cannot be used to compare languages and are unable to provide a holistic measure of language production.

The Pairwise Variability Index (PVI) was developed by Low, et al. [11] and Grabe and Low [5] in accordance with Dauer's [4] proposal that rhythm involves vowel duration and syllable structure. Ramus, et al. [13] also developed interval measures (⊿C, ⊿V, and %V) which are based on a similar concept. The advantage of the use of these measures was that they were able to provide an acoustic correlate of rhythm which could identify languages that have traditionally been categorized as syllable and stress timed languages.

Kinoshita [9, 10] determined that the PVI correlated more highly with the ratio measures and was also able to distinguish between beginning and advanced learners, and concluded that this measure was the more useful of the two for measuring second language acquisition. The PVI is beginning to be used in SLA research e.g. [6, 16].

Different rhythmic categories have their origins in perception [1]. Thus, in order to evaluate the validity of the Index it would be useful to compare its results with the perceptual judgments of listeners. White, et al. [17] has demonstrated that native English listeners are able to discriminate between languages with different acoustic measures. However, his results had not been duplicated with second language learners.

This paper reports a study which intended to discover if there was a relationship between the PVI measures and native speaker ratings of rhythm. Such a relationship would validate the use of the PVI index in second language acquisition research.

The research questions then, were:

1) to what extent do raters evaluate speech rhythm in common, and
2) are native speaker ratings of speech rhythm related to the PVI?

## 2. METHOD

In order to determine the validity of the use of the Pairwise Variability Index, the production of second language learners was evaluated by native speakers and then these evaluations were correlated with the PVI measures (nPVIv and nPVIc).

### 2.1. Participants

16 Korean Learners of Japanese as a second language took part in this study. They included students enrolled in a Korean university as Japanese majors and overseas students enrolled in Japanese universities[1].

5 native speakers of Japanese rated the production of these learners. They had all completed Master's degrees in Japanese Education, majoring in pronunciation and communication, and therefore could be considered to be experts.

### 2.2. Materials

A Japanese translation of Aesop's tale 'The North Wind and the Sun' was used as the speech sample.

10 Native speaker samples of the text were taken from a Japanese language speech corpus [15].

### 2.3. Procedure

The procedure was as follows. Readings of the fable were recorded with a Sony PCM (PCM-D50) recorder in a quiet environment. After practicing, the participants recorded their own versions until they were satisfied with the results.

The recordings from both the Korean participants and the Japanese corpus were then cut into 8 sentences using the acoustic analysis software package Praat ver. 4.4.34 [2]. The segmentation procedures followed Grabe and Low [5], and identified the vocalic and intervocalic duration in each sentence. The resulting vocalic (nPVIv) and intervocalic (nPVIc) indexes were calculated for the 208 sentences.

Evaluations of the Korean non-native readings were carried out by placing the sentences in random order and recording them onto an audio CD. The raters then listened to each sentence using headphones and rated them on a four point Likert scale according to the extent of their agreement that 'this speaker has good speech rhythm'.

### 2.4. Analysis

In order to answer the first research question, oblique (non-orthogonal) exploratory factor analyses (direct oblimin, delta = 0) were carried out on the ratings for each of the eight sentences. In addition Cronbach alphas were calculated for each of the resulting factors. Raters not contributing to this internal consistency were removed from further analysis involving that factor.

The second research question attempted to determine if variance in the ratings could be explained by the PVI measures. Z-scores were calculated for the Korean participants' data based on the Japanese native speaker data mean and standard variation, providing a baseline. The resulting z-scores (nPVIv and nPVIc) and averaged ratings were then Winsorized with 20% trimming [18] to account for the small sample size and reduce any effects of outliers. Any possible relationships between the data were then confirmed using line of fit on SPSS. The fit which explained the most variance in the ratings is reported here.

Significance was set at $p < .05$[2].

## 3. RESULTS

### 3.1. Rater internal consistency

The results of the factor analysis indicated that the rhythm ratings were not unidimensional across raters with the ratings on sentences 1, 5, 6, 7, and 8 being represented by at least 2 underlying factors. The percentage of the data's variance explained by each factor is shown in table 1.

**Table 1:** The number of factors and the % variance explained underlying the ratings of each sentence.

| Sentence | % Variance Explained | |
|---|---|---|
| | Factor 1 | Factor 2 |
| 1 | 42.729 | 29.372 |
| 2 | 61.621 | - |
| 3 | 71.549 | - |
| 4 | 70.605 | - |
| 5 | 51.411 | 26.251 |
| 6 | 56.007 | 24.906 |
| 7 | 51.885 | 24.641 |
| 8 | 57.309 | 23.981 |

Further analysis examined the internal consistency between the raters for each factor. The results, shown on table 2, demonstrated that in every case, ratings associated with the first factor reached an internal consistency ($\alpha > .7$) acceptable for exploratory research such as this. When there was a second factor, the internal consistency failed to reach the required standard.

Based on these results, only the evaluations associated with the first factor were used for the analysis in part 2.

**Table 2:** Internal consistency and the number of raters for each factor and sentence

| Sentence | Factor 1 | | Factor 2 | |
|---|---|---|---|---|
| | No. of Raters | Alpha | No. of Raters | Alpha |
| 1 | 3 | 0.707 | 3 | 0.499 |
| 2 | 3 | 0.913 | - | - |
| 3 | 3 | 0.937 | - | - |
| 4 | 5 | 0.895 | - | - |
| 5 | 4 | 0.770 | 3 | 0.697 |
| 6 | 3 | 0.894 | 2 | 0.658 |
| 7 | 3 | 0.844 | 2 | -* |
| 8 | 3 | 0.842 | 3 | -* |

### 3.2. Rhythm ratings and the PVI measures

The results of the line of fit analysis demonstrated that significant variance in the ratings was explained by variability in vowel lengths rather than consonant lengths (Table 3). In the single case where a significant amount of variance was explained by nPVIc (sentence 3, $R^2 = .305$), a step-wise multiple regression demonstrated that the

correlation between the ratings and nPVIv could account for a relationship between nPVIv and nPVIc present in sentence 3.

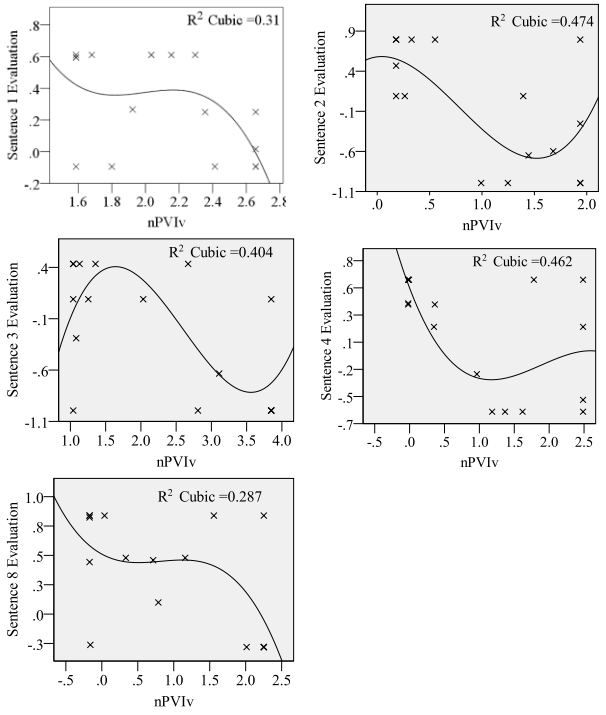
**Table 3:** The Line of Fit and % of variance in ratings explained ($R^2$) by nPVIv.

| Sentence | Line of Fit | $R^2$ |
|---|---|---|
| 1 | Cubic | 0.310 |
| 2 | Cubic | 0.474 |
| 3 | Cubic | 0.404 |
| 4 | Cubic | 0.462 |
| 8 | Cubic | 0.287 |

*p < .05

Significant lines of fit were found between the ratings and nPVIv for five (Sentence 1, 2, 3, 4 and 8) of the eight sentences. The scatter plots shown in Figure 1 demonstrate that, for the greater part, ratings were higher when nPVIv was closer to the native speaker mean (z = 0). In every case, the best line of fit was cubic.

**Figure 1:** Scatter plots and significant lines of fit for sentence ratings 1, 2, 3, 4 and 8 and nPVIv (z-score).




In order to explain why the PVI accounted for rater variation in some sentences and not others, a comparison between the participants' speech production and that of the native speaker corpus was undertaken. Results of the t-test comparisons are shown on Table 4. There are no significant differences between the nPVIv of the Korean participants and Japanese native speakers for Sentences 5, 6, 7, and 8.

**Table 4:** T-test comparisons of the speech production between Japanese native speakers and the Korean participants.

| Sentence | nPVIv | | | nPVIc | | |
|---|---|---|---|---|---|---|
| | t | df | P | T | df | p |
| 1 | 4.755 | 24 | .000 | .090 | 24 | .929 |
| 2 | 2.683 | 24 | .013 | 4.019 | 24 | .001 |
| 3 | 3.605 | 24 | .001 | 3.988 | 24 | .001 |
| 4 | 2.154 | 24 | .042 | 1.462 | 24 | .157 |
| 5 | .056 | 24 | .956 | .179 | 24 | .860 |
| 6 | 1.261 | 24 | .219 | 4.398 | 24 | .000 |
| 7 | .155 | 24 | .878 | .455 | 24 | .653 |
| 8 | 1.676 | 24 | .107 | .471 | 24 | .642 |

## 4. DISCUSSION

The results have clearly demonstrated that the raters do not always base their ratings on a single underlying construct of speech rhythm. Firstly, there appears to be intra-rater variability. This is clear from the multiple factor solutions obtained from the factor analysis. Five of the eight sentences produced two factor solutions indicating at least two underlying constructs associated with speech rhythm evaluation. Further evidence for this comes from the fact that not all raters contributed to the internal consistency of the ratings. In fact, one of the five raters was only included as rating in common with the other raters for one of the eight sentences.

The implications of this are that different raters select different aspects of speech rhythm when they are evaluating identical speech samples.

Secondly it appears that there is Intra-rater variability across the ratings of speech samples. This is evident in the variability of individual's judgments across sentences. Ratings of three sentences were not significantly related to the PVI measures, whereas the other five were. It appears a single rater will change the criteria used to evaluate speech rhythm depending on the speech sample.

Regardless of this variability, there also appears to be commonality underlying the speech rhythm ratings of these expert native speakers. All sentences produced Cronbach alphas in the 'acceptable' range of .71- .94.

The second research question examined the extent to which PVI measures and expert native speaker ratings were related. The findings show that there is a relationship between the ratings and variability in pairwise vowel lengths (nPVIv). The relationship also appears to be consistently cubic in nature which accounts for quite large (.287 - .474) amounts of variance in the ratings.

This result was not replicated in all the data, with three sentences demonstrating no significant relationship between the two variables. This discrepancy can be explained by the fact that these three sentences also showed no significant differences between the native speakers and the non-native participants in the nPVIv measures. This indicates that when non-native speakers attain native-speaker levels, the PVI measures cease to contribute to variability in the ratings.

No relationship was found between the pairwise intervocalic lengths (nPVIc). This is in spite of the fact that the non-native participants differed significantly from the native speakers in their nPVIc scores on three separate sentences.

One possible reason for this result could be the low ratio of closed syllables (geminate consonants and nasal consonants) in Japanese, 8% according to Ohtake [12]. However, that does not seem to be the case for this speech. Sentences 3, 5, and 7 all included 2 - 3 closed syllables. This was comparable to the occurrence of long vowels.

## 5. CONCLUSION

In conclusion, the following can be inferred from this study:

1) Although there is Intra- and Inter rater variability there also appears to be a common factor underlying native speaker judgments in ratings of speech rhythm,
2) There is a consistent cubic relationship between such ratings and nPVIv. This indicates clearly that native speakers use variability in vowel length, among other things, when judging non-native speaker speech rhythms.

These results suggest that the PVI is a useful measure of non-native speech rhythm and can be useful in second language acquisition research.

Future investigations need to examine further possible acoustic rhythm correlates which could account for variance in the ratings as half of the variability still remains unexplained.

## 6. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh University Press.

[2] Boersma, P., Weenink, D. 2006. Praat: Doing Phonetics by Computer. *http://www.fon.hum.uva.nl/praat*

[3] Cho, M. 2004. Rhythm typology of Korean speech, *Cognitive Processing* 5, 249-253.

[4] Dauer, R.M. 1983. Stress-timing and syllable-timing re-analyzed. *Journal of Phonetics* 11, 51-62.

[5] Grabe, E., Low, E. 2002. Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology* 7, 515-546.

[6] Grenon, I., White, L. 2008. Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In Chan, H., Jacob, H., Kapia, E. (eds.), *BUCLD 32: Proceedings of the 32nd Annual Boston University Conference on Language Development*. Somerville, Massachusetts: Cascadilla Press, 155-166.

[7] Kashima, T. 2001. *Nihongo Hatsuon Kyoiku eno Oyo wo Mezashita Aatarashi Rizumu Tan'i no Onseteki Jitsugen ni Kansuru Kiso Kenkyu*. Hese 9 Nendo Kagaku Kenkyuhi Hojokin (Kiban Kenkyu(C)(2)) Kenkyu Seka Hokokusho.

[8] Kinoshita, N. 2009. Reconsidering the rhythm of Korean, *Meikai Journal, Faculty of Languages and Cultures* 21, 169-177.

[9] Kinoshita, N. 2011. *The Acquisition and Teaching of Japanese Rhythm*. Waseda University Press.

[10] Kinoshita, N., Sheppard, C. 2008. The acquisition of lexical rhythm and duration by Japanese second language learners. *Laboratory Phonology* 11, 79-80.

[11] Low, E.L., Grabe, E., Nolan, F. 2000. Quantitative characterizations of speech rhythm: 'Syllable-timing' in Singapore English. *Language and Speech* 43, 377-401.

[12] Ohtake, T. 1990. Gengo no Rizumu to Onsetsukozo. *Denshi Joho Tsushin Gakkaishi* 89, 55-61.

[13] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal, *Cognition* 73, 265-292.

[14] Sukegawa, Y. 1993. Bogobetsu ni Mita Hatsuon no Keiko: Anketo no Kekka kara. *Nihongo Onse to Nihongo Kyoiku*, Monbusho Jutenryoiki Kenkyu., Nihongo Onse ni okeru Inritsuteki Tokucho no Jittai to Sono Kyoiku ni kansuru Sogoteki Kenkyu, 187-222.

[15] Suzuki, H. 2002. *Tsukuba University Tozaigengo Bunka no Ruikeron. Tokubetsu Purojekuto Kenkyu Kenkyuseka Hokokusyo bessatsu. Tagengoonse Kopasu*. Hese 13 Nendo. Tsukuba University Tozaigengo Bunka no Ruikeron Tokubetsu Purojekuto Kenkyu Soshiki, 36.

[16] White, L., Mattys, S.L. 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35, 501-522.

[17] White, L., Mattys, S.L., Series, L., Gage, S. 2007. Rhythm metrics predict rhythmic discrimination. *Proceedings of the 16th International Congress of the Phonetic Sciences*, 1009-1012.

[18] Wilcox, R. 2010. *Fundamentals of Modern Statistical Methods* (2nd ed.). New York, NY: Springer.

[1] Although there some (e.g. Cho [3]) who suggest that Japanese rhythm and Korean rhythm are similar, Kinoshita [8] determined that the rhythms of English, Japanese and Korean rhythm, as measured by the PVI, were significantly different. Kinoshita [9, 10] also determined that Korean speakers have difficulty learning Japanese rhythm. This finding is also born out in this study.

[2] This corresponds to $r > .468$, or $r^2 > .219$ for this data set.